Argonne NATIONAL LABORATORY

Early Experience with I/O on Mira

Venkat Vishwanath

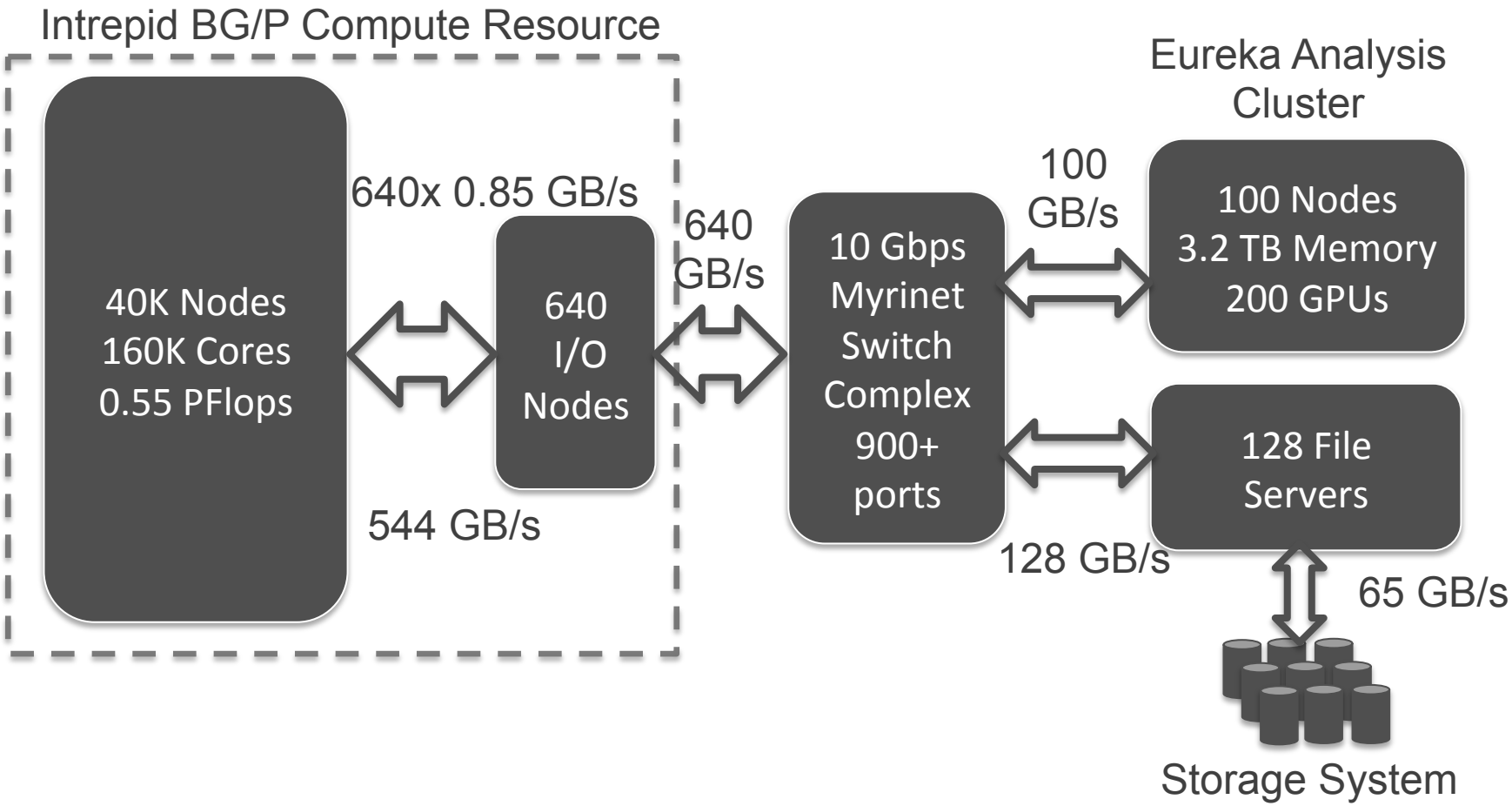
Steve Crusan and Kevin Harms

Argonne National Laboratory

venkat@anl.gov

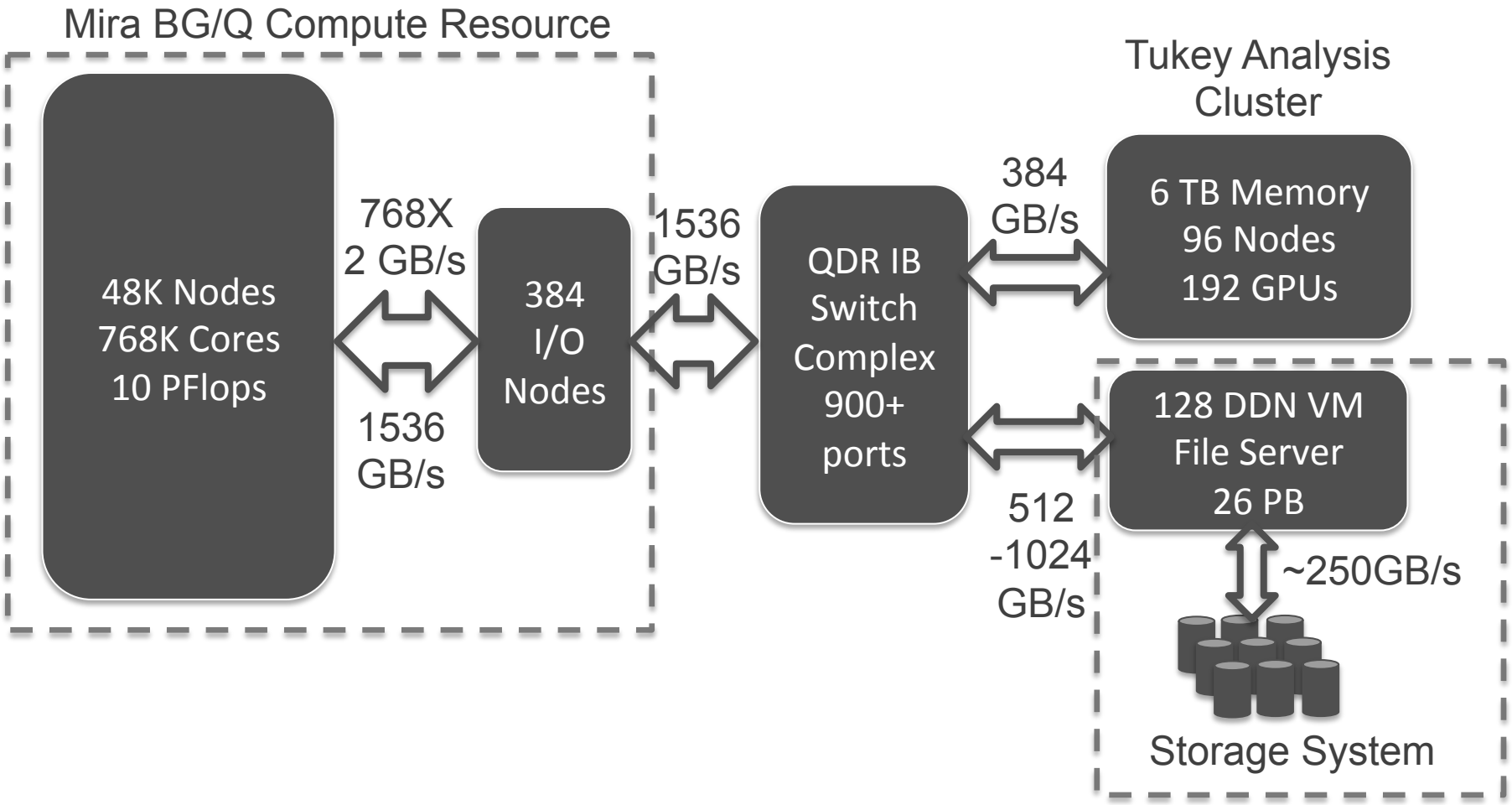
Disclaimer: Mira I/O subsystem and filesystem is still work in progress. The configuration and performance will change, and we will keep users informed of these.

ALCF-1 I/O Infrastructure



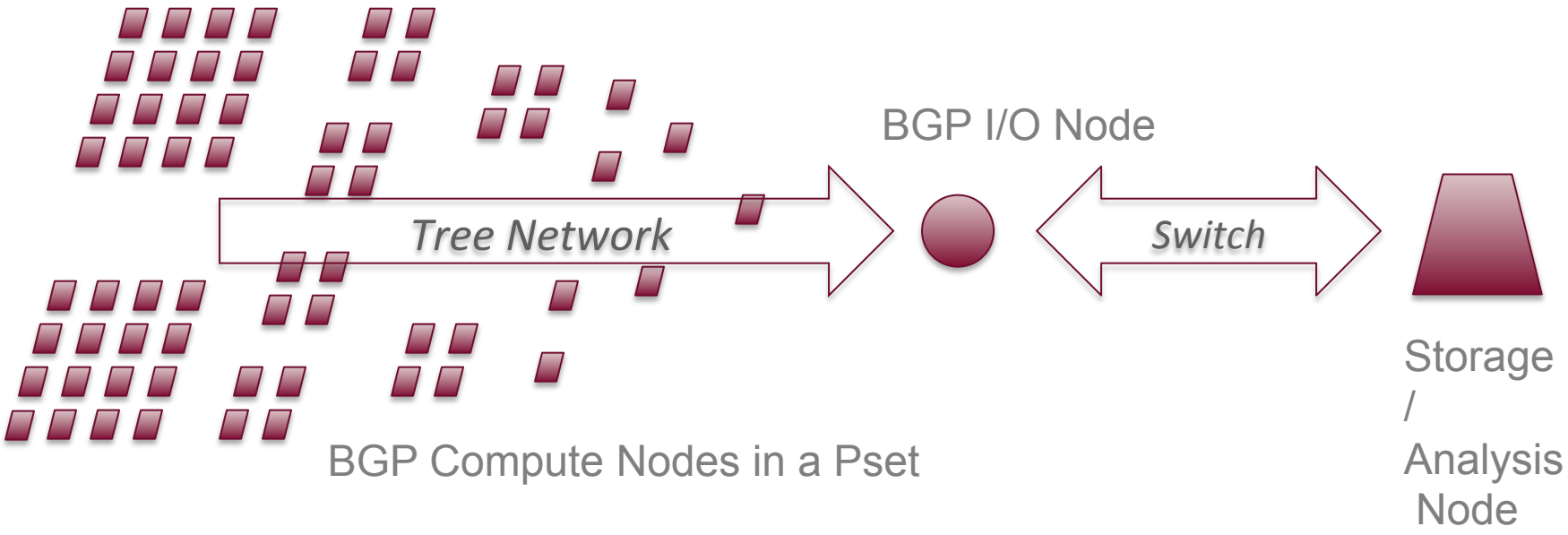
ALCF uses the GPFS filesystem for production I/O

ALCF-2 I/O Infrastructure



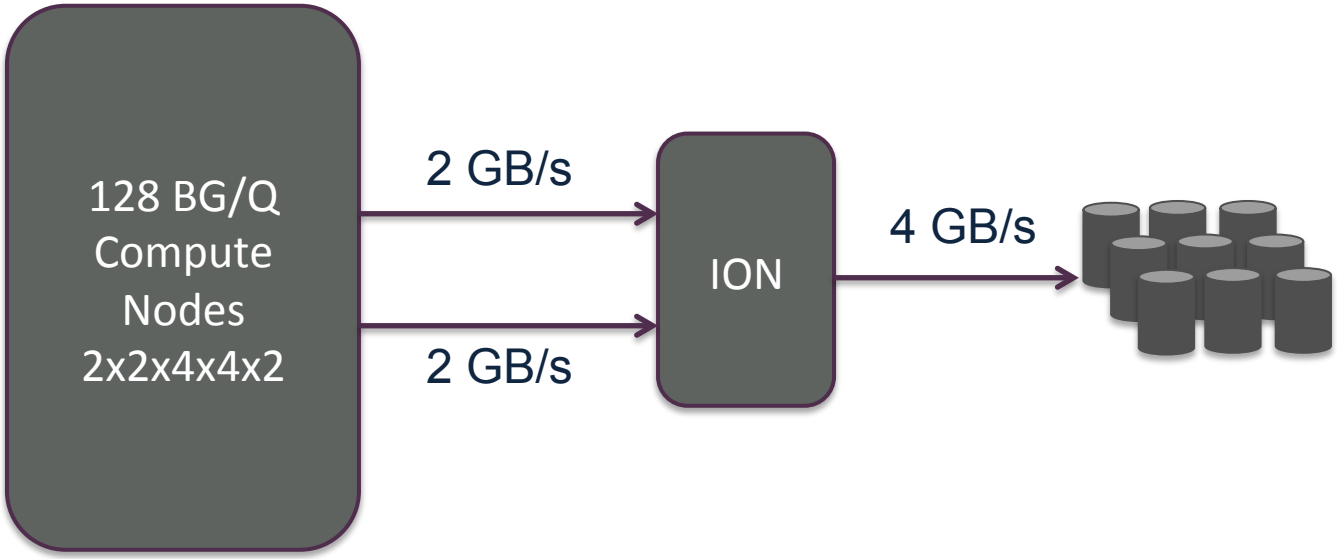
Storage System consists of 16 DDN SFA12KE “couplets” with each couplet consisting of 560 x 3TB HDD, 32 x 200GB SSD, 8 Virtual Machines (VM) with two QDR IB ports per VM

I/O Subsystem of BG/P



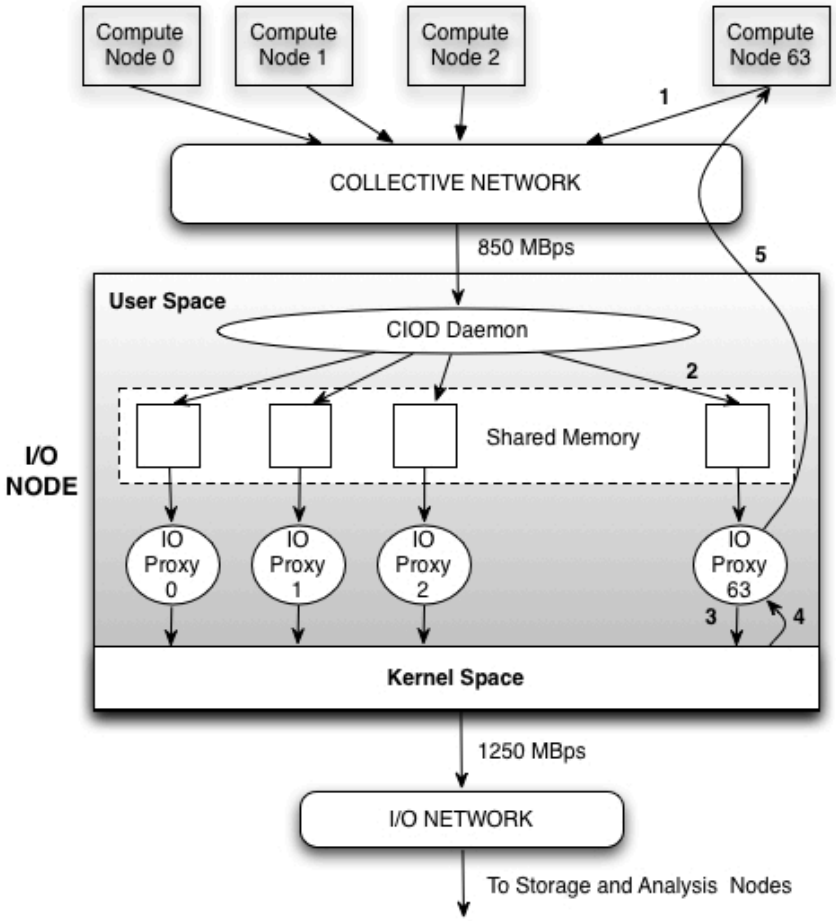
I/O forwarding is a paradigm that attempts to bridge the increasing performance and scalability gap between the compute and I/O components of leadership-class machines to meet the requirements of data-intensive applications by shipping I/O calls from compute nodes to dedicated I/O nodes.

I/O Subsystem of BG/Q- Simplified View



- For every 128 compute nodes, we have two links to the I/O Node on the BG/Q Torus
- On the ION, we have a QDR Infiniband connected on a PCI-e Gen2 Slot

Control and I/O Daemon (CIOD) - I/O Forwarding mechanism for BG/P and BG/Q



- CIOD is the I/O forwarding infrastructure provided by IBM for the Blue Gene / P
- For each compute node, we have a dedicated I/O proxy process to handle the associated I/O operations
- On BG/Q, we have a **thread based** implementation instead of process-based

Intrepid vs Mira : I/O Comparison

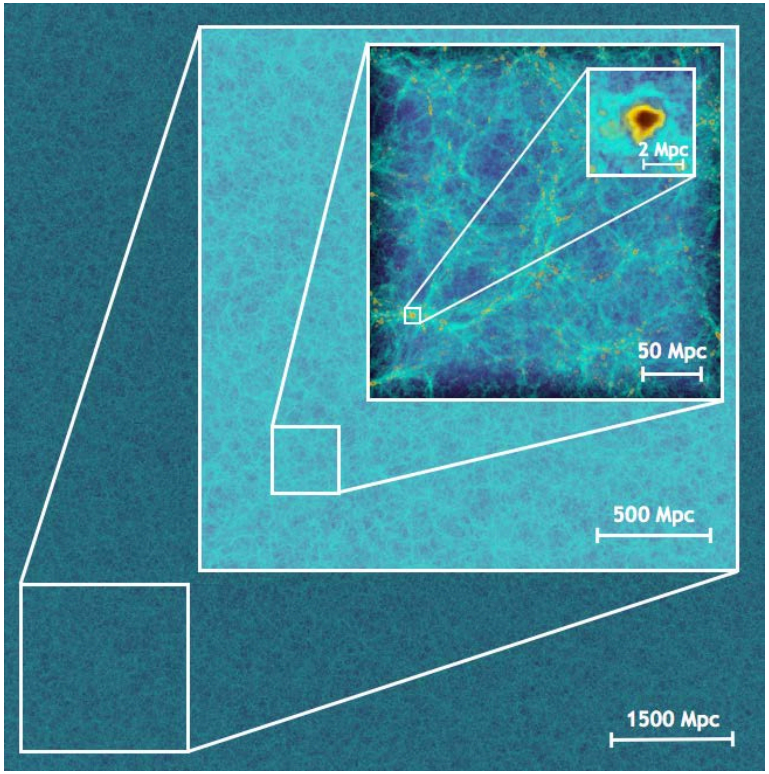
	Intrepid BG/P	Mira BG/Q	Increase
Flops	0.557 PF	10 PF	20X
Cores	160K	768K	~5X
ION / CN ratio	1 / 64	1 / 128	0.5X
IONs per Rack	16	8	0.5X
Total Storage	6 PB	35 PB	6X
I/O Throughput	62GB/s	240 GB/s Peak	4X
User Quota?	None	Quotas enforced!!	-
I/O per Flops (%)	0.01	0.0024	**0.24X**

I/O interfaces and Libraries on Mira

- Interfaces
  - POSIX
  - MPI-IO
- Higher-level Libraries
  - HDF5
  - Netcdf4
  - PnetCDF
- Early I/O Performance on Mira
  - Shared File Write ~70 GB/s
  - Shared File Read ~180 GB/s

Early Experience Scaling I/O for HACC

- Hybrid/Hardware Accelerated Computational Cosmology code is lead by Salman Habib at Argonne
- Members include Katrin Heitmann, Hal Finkel, Adrian Pope, Vitali Morozov
- Particle-in-Cell code
- Achieved ~14 PF on Sequoia and 7 PF on Mira
- On Mira, HACC uses 1-10 Trillion particles



Zoom-in visualization of the density field illustrating the global spatial dynamic range of the simulation -- approximately a million-to-one

Early Experience Scaling I/O for HACC

- Typical Checkpoint/Restart dataset is **~100-400 TB**
- Analysis output, written more frequently, is **~1TB-10TB**
- Step 1:
  - Default I/O mechanism using single shared file with MPI-IO
  - Achieved ~15 GB/s
- Step 2:
  - File per rank
  - Too many files at 32K Nodes (262K files with 8RPN)
- Step 3:
  - An I/O mechanism writing 1 File per ION
  - Topology-aware I/O to reduce network contention
  - Achieves up to **160 GB/s** (~10 X improvement)
- Written and read **~10 PB** of data on Mira (and counting)

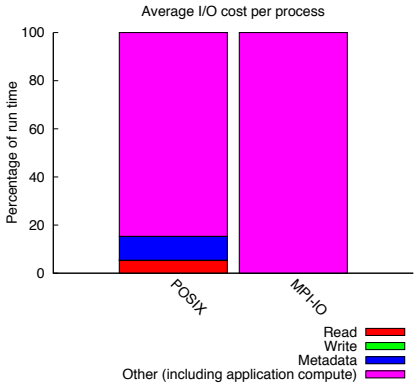
Few Tips and Tricks for I/O on BG/Q

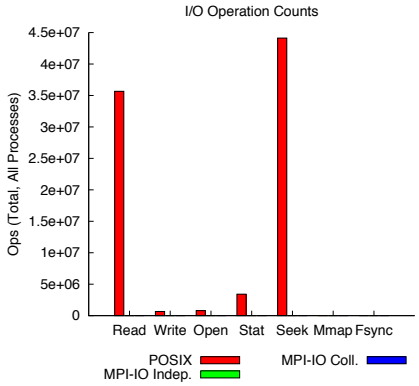
- Pre-create the files
  - Increases shared file write performance from 70GB/s to 130GB/s
- Single shared file vs file per process (MPI rank)
- Pre-allocate files
- HDF5 Datatype
  - Instead of Native, use Big Endian(eg. H5T_IEEE_F64BE)
- MPI Datatypes
- Collective I/O
- MPI-IO BgLockless Environment variable
  - Prefix file with bglockless:/
  - BGLOCKLESSMPIO_F_TYPE=0x47504653

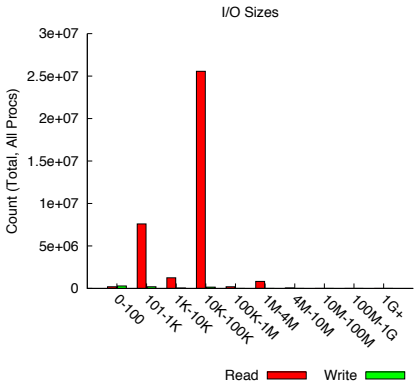
Darshan - An I/O Characterization Tool

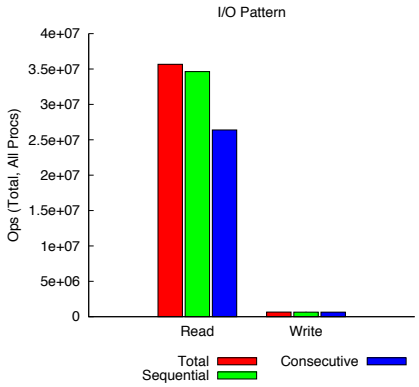
- An open-source tool developed for statistical profiling of I/O
- Designed to be lightweight and low overhead
  - Finite memory allocation for statistics (about 2MB) done during MPI_Init
  - Overhead of 1-2% total to record I/O calls
  - Variation of I/O is typically around 4-5%
  - Darshan does not create detailed function call traces
- No source modifications
  - Uses PMPI interfaces to intercept MPI calls
  - Use ld wrapping to intercept POSIX calls
  - Can use dynamic linking with LD_PRELOAD instead
- Stores results in single compressed log file
- http://www.mcs.anl.gov/darshan

Darshan on BG/Q









Most Common Access Sizes

access size	count
65536	23066526
800	2378602
304	2172322
400	2029354

- logs:
/gpfs/vesta-fs0/logs/darshan/vesta
/gpfs/mira-fs0/logs/darshan/mira
- bins:
/soft/perftools/darshan/darshan/bin
- wrappers:
/soft/perftools/darshan/darshan/wrappers/<wrapper>
- export PATH=/soft/perftools/darshan/darshan/wrappers/xl:${PATH}

Questions